

CLAIMS:

1. A method of identifying a set of biologically-active DNA-binding sites for a protein of interest in the genome of a cell, the method comprising
  - 5 (i) identifying a set of regions of genomic DNA to which the protein of interest is bound in the cell;
  - (ii) identifying candidate DNA-binding sites in the identified regions of genomic DNA, wherein a candidate DNA-binding site comprises a sequence corresponding to a DNA-sequence motif for the protein of interest;
  - 10 (iii) determining if the candidate DNA-binding sites are conserved in an equivalent genomic region in one or more species different from the species from which the cell is obtained, wherein a candidate DNA-binding site that is conserved in at least one of the different species is
  - 15 a biologically-active DNA-binding site.
2. The method of claim 1, wherein step (i) further comprises identifying a DNA-sequence motif for the protein from the set of regions of genomic DNA.
- 20 3. The method of claim 2, wherein the DNA-sequence motif is enriched by a statistically-significant amount in the set of regions of genomic DNA relative to a suitable control.
- 25 4. The method of claim 3, wherein the suitable control comprises a set of genomic regions which are not bound by the protein of interest in the cell.
5. The method of claim 3, wherein the suitable control comprises a set of randomly selected genomic regions.
- 30 6. The method of claim 3, wherein the suitable control comprises a set of randomly generated sequences.

7. The method of claim 3, wherein the suitable control comprises a set of genomic regions which are bound by a mutant form of the protein of interest in the cell.
- 5 8. The method of claim 1, wherein the regions of genomic DNA comprise promoter regions.
9. The method of claim 1, wherein the regions of genomic DNA have a length  
10 from about between 50 bp to about 10 kb.
10. The method of claim 1, wherein step (i) comprises performing genome-wide location analysis (GWLA) of the protein of interest.
- 15 11. The method of claim 10, wherein GWLA comprises chromatin-immunoprecipitation (ChIP) and subsequent analysis on DNA microarray (ChIP-chip).
12. The method of claim 2, wherein a candidate DNA-binding site is conserved  
20 if the equivalent genomic region in at least one different species comprises a nucleic acid sequence that matches the DNA-sequence motif for the protein of interest.
13. The method of claim 2, wherein the DNA-sequence motif is identified using  
25 at least one algorithm.
14. The method of claim 13, wherein the algorithm is selected from the group consisting of AlignACE, MEME, MDscan, the Kellis Method, Mogul, Verbumculus, YMF, BioProspector, Motif Sampler and SUPERPOSITION.
- 30 15. The method of claim 2, wherein the DNA-sequence motif is identified using a combination of algorithms.

16. The method of claim 1, wherein the candidate DNA-binding site is less than 20 bp in length.
- 5 17. The method of claim 1, wherein the DNA-sequence motif is degenerate in at least one position.
18. The method of claim 1, wherein one or more of the different species is classified in the same genus as the cell.
- 10 19. The method of claim 1, wherein step (iii) comprises determining if the candidate DNA-binding sites are conserved in equivalent genomic regions in two or more different species
- 15 20. The method of claim 1, wherein the protein of interest is a transcriptional regulator.
21. The method of claim 1, wherein the protein of interest comprises a DNA binding domain.
- 20 22. The method of claim 1, wherein the protein of interest does not comprise a DNA-binding domain.
- 25 23. The method of claim 21 or 22, wherein the DNA-binding domain is selected from the group consisting of zinc finger, winged-helix, leucine zipper, homeodomain and helix-loop-helix (HLH).
24. The method of claim 1, wherein the set of biologically-active DNA-binding sites comprises one or more biologically-active DNA-binding sites.
- 30 25. The method of claim 1, wherein the set of biologically-active DNA-binding sites comprises 10 or more biologically-active binding sites.

26. The method of claim 1, wherein two regions of genomic DNA are equivalent if they both comprise a sequence of at least one orthologous gene.
- 5 27. The method of claim 1, wherein two regions of genomic DNA, each comprising an intergenic region which is flanked by a first and a second open reading frame (ORF) in their respective genomes, are said to be equivalent if (i) the first ORF in the two regions are orthologous ORFs and (ii) if the second ORFs in the two regions are orthologous ORFs.
- 10 28. The method of claim 1, wherein the cell is an eukaryotic cell.
29. The method of claim 28, wherein the cell is a stem cell.
- 15 30. The method of claim 28, wherein the cell is a mammalian cell.
31. The method of claim 30, wherein the cell is a human cell.
32. The method of claim 1, wherein the cell is a primary cell.
- 20 33. The method of claim 31, wherein the cell is derived from a tissue biopsy.
34. The method of claim 33, wherein the tissue biopsy is isolated from a subject afflicted with a disorder.
- 25 35. The method of claim 1, wherein the cell is a single-cell organism.
36. A method of identifying an agent which alters the set of biologically-active DNA-binding sites for a protein of interest in the genome of a cell, the method comprising
- 30 (i) contacting an experimental cell with a candidate agent;
- (ii) identifying a set of biologically-active DNA-binding sites for a

protein of interest in the genome of the cell of step (i) according to the method of claim 2, thereby generating an experimental set of biologically-active DNA-binding sites;

(iii) comparing

- 5           (1) the experimental set of biologically-active DNA-binding sites to  
            (2) a control set of biologically-active DNA-binding sites for the protein of interest;  
            wherein a candidate agent is identified if the experimental set and the control set differ.

10

37. The method of claim 36, wherein the control set is derived from a control cell that is not contacted with the candidate agent.

15

38. A method of identifying a pathway that is transcriptionally regulated by a protein of interest in a cell, the method comprising

(i) identifying a set of biologically-active DNA-binding sites for a protein of interest in the genome of the cell according to the method of claim 2; and

20

(ii) identifying at least two candidate genes likely to be regulated by binding of the protein of interest to the set of biologically-active DNA-binding sites identified in (i);

wherein a pathway that is transcriptionally regulated by the protein of interest is identified if at least two candidate genes are members of the same pathway.

25

39. The method of claim 38, further comprising modulating the pathway that is transcriptionally regulated by the protein of interest, by exposing a cell to an agent or condition which alters a set of biologically-active DNA-binding sites for the protein of interest.

30

40. The method of claim 38, wherein the pathway is a biochemical pathway.

41. The method of claim 38, wherein the pathway is a gene expression pathway.
42. The method of claim 38, wherein the pathway is a regulatory pathway.
- 5 43. The method of claim 38, wherein a candidate gene is likely regulated by the protein of interest if the promoter for the candidate gene comprises at least one biologically-active DNA-binding site.
- 10 44. The method of claim 43, wherein the promoter region of a candidate gene comprises from about 3kb 5' to about 1kb 3' of the transcription initiation site.
45. A method of identifying two sets of conditions in which a protein of interest differentially binds to the genome of a cell, the method comprising:
- 15 (i) identifying a first set of biologically-active DNA-binding sites for the protein of interest in the genome of a cell according to the method of claim 1, wherein the cell is exposed to a first set of conditions;
- (ii) identifying a second set of biologically-active DNA-binding sites for the protein of interest in the genome of a cell according to the method
- 20 of claim 1, wherein the cell is exposed to a second set of conditions; and
- (iii) comparing the first set of biologically-active DNA-binding sites to the second set of biologically-active DNA-binding sites and determining if the two sets differ.
- 25 46. A method of identifying a property of a gene product of a gene of interest that correlates with the binding activity of a polypeptide encoded by the gene of interest to the genome of a cell, the method comprising
- (i) identifying two sets of conditions in which a protein of interest
- 30 differentially binds to the genome of the cell according to the method of claim 44;
- (ii) determining a property of a gene product of the gene of interest in (a)

- a cell exposed to the first set of conditions; and in (b) a cell exposed to the second set of conditions; and
- (iii) determining if at least one property of the gene product differs in the two cells of step (ii),
- 5 thereby identifying a property that correlates with the binding activity of a gene of interest to the genome of a cell.
47. A method of identifying a property of a gene product of a gene of interest that correlates with the binding activity of a polypeptide encoded by the gene of interest to the genome of a cell, the method comprising
- 10 (i) identifying an agent which alters the set of biologically-active DNA-binding sites for a protein of interest in the genome of a cell according to the method of claim 36;
- (ii) determining a property of a gene product of the gene of interest in (a)
- 15 a cell contacted with the agent; and in (b) a cell not contacted with the agent; and
- (iii) determining if at least one property of the gene product differs in the two cells of step (ii),
- thereby identifying a property that correlates with the binding activity of a
- 20 gene of interest to the genome of a cell.
48. The method of claim 46 or 47, wherein the property is selected from the group consisting of a protein modification, expression level, enzymatic activity and intracellular localization.
- 25
49. The method of claim 46 or 47, wherein the expression product is an mRNA.
50. The method of claim 46 or 47, wherein the expression product is a polypeptide.
- 30
51. The method of claim 46 or 47, wherein the property comprises the expression level of the gene product.

52. The method of claim 46 or 47, wherein the property comprises the subcellular localization of the gene product.
- 5 53. The method of claim 46 or 47, wherein the property comprises the phosphorylation state of gene product.
54. The method of claim 46 or 47, wherein the property comprises the molecular weight of the gene product.
- 10 55. The method of claim 46 or 47, wherein the property comprises the isoelectric point of the gene product.
56. The method of claim 46 or 47, wherein the property comprises the nucleic acid sequence or the amino acid sequence of the gene product.
- 15 57. The method of claim 46 or 47, wherein the property comprises the physical association of the protein of interest with another polypeptide.
- 20 58. The method of claim 46 or 47, wherein the property comprises an enzymatic activity of a polypeptide gene product.
59. The method of claim 46 or 47, wherein the property comprises the oligomeric state of a polypeptide gene product.
- 25 60. A method of identifying two cell genotypes in which a protein of interest differentially binds to the genome of a cell, the method comprising:
- (i) identifying a first set of biologically-active DNA-binding sites for the protein of interest in the genome of a cell of a first genotype;
- 30

- (ii) identifying a second set of biologically-active DNA-binding sites for the protein of interest in the genome of a cell of a second genotype;
- (iii) comparing the first set of biologically-active DNA-binding sites to the second set of biologically-active DNA-binding sites and determining if the two sets differ.

5